Geometry of the Energy Landscape for a Protein Folding on the Ribosome

David S. Tourigny*

MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK

(Dated: April 29, 2013)

Abstract

Energy landscape theory describes how a full-length protein can attain its native fold by sampling only a tiny fraction of all possible structures. Although protein folding is now understood to be concomitant with synthesis on the ribosome, there have been few attempts to modify energy landscape theory by accounting for cotranslational folding. Here we provide a model for cotranslational folding that leads to a natural definition of a nested energy landscape. By applying concepts drawn from submanifold differential geometry, the physics of protein folding on the ribosome can be explored in a quantitative manner and conditions on the nested energy landscapes for a good cotranslational folder are derived.

PACS numbers: 87.15.-v; 02.40.-k

A fundamental problem in molecular biology is explaining how the three-dimensional structure of a protein is encoded within its amino acid sequence. Inside the cell, proteins are synthesised on the ribosome by sequential addition of residues to an elongating polypeptide chain during a process called translation [1]. Translation accounts for the conversion of genetic information to the primary sequence of a protein, but knowledge of how the molecule then folds into a functional state is central to our understanding of the natural world. Energy landscape theory provides a mechanism whereby the existence of intermediate structures, each associated with a free energy cost, enable the folding pathway of a protein to be mapped on a multidimensional potential energy landscape. Assuming the global shape of the energy landscape for a good folder resembles a funnel means that only a small fraction of all possible structures need to be sampled before the protein attains its native fold [2–4].

Nascent proteins can begin to fold whilst they are still bound to a translating ribosome [5, 6]. During cotranslational folding, the conformational space available to a protein increases incrementally with addition of residues to the polypeptide chain. This can enhance folding yields [7], provide an additional level of quality control [8, 9], and allow access to folding pathways different from those available to a full-length protein [10, 11]. Many studies (e.g. [12–14]) have highlighted a relationship between folding timescales and the delay before amino acid addition (τ_A), which can be modulated by codon usage and controlled by the translational apparatus. Some research groups have also developed a theoretical understanding of how protein folding is affected by varying translation rates τ_A [15, 16], but so far there has not been a satisfactory attempt to modify energy landscape theory by accounting for cotranslational folding.

Taking a geometric approach this problem could reveal a relationship between τ_A and curvatures of the energy landscape. Moreover, it should account recursively for the properties of the energy landscape at different chain lengths. In the past, pseudo-Riemannian geometry has been successfully applied to folding of full-length proteins [17, 18], but the questions we wish to address here lead to the development of a quite different theory specific to cotranslational folding. The result is an analytical set of conditions that must be satisfied by a group of *nested energy landscapes* for any protein folding cotranslationally on the ribosome.

We begin our treatment as in [17, 18], by considering an enlarged $(N+2)$-dimensional configuration space with coordinates $q^0, q^1, ..., q^N, q^{N+1}$, where $q^1, ..., q^N$ are the Lagrangian

coordinates of a polypeptide of length n. Denote a potential energy function on this space by $V_n = V_n(q^1, ..., q^N)$. In keeping with the assumptions of [15, 16], the transition from nascent chain length n to $(n+1)$ is instantaneous relative to the times that the ribosome spends at either of these chain lengths. Consequently, addition of a new amino acid to the polypeptide results in a discrete jump of value $V_{n+1} - V_n$ in the potential energy associated with any given conformation of the nascent chain.

The configuration spaces of the nth and $(n-1)$th states can be represented as manifolds M and $\bar{M}$ respectively. We endow the manifold M with an Eisenhart metric [19] whose arc length is

$$ds^2 = \delta_{ij} dq^i dq^j - 2V_n (dq^0)^2 + 2dq^0 dq^{N+1} \ , \tag{1}$$

where the indices i, j run from 1 to N. We want the projections of the geodesics of this Eisenhart metric to be the natural motions of the Hamiltonian system and therefore the folding trajectories of a protein in the nth state. These trajectories, parameterised by the time coordinate $q_0 = t$, are obtained by taking $ds^2 = dt^2$ on the physical geodesics and imposing the integral condition

$$q^{N+1} = \frac{1}{2} t + c_0 - \int_0^t \left[\delta_{ij} \dot{q}^i \dot{q}^j - V_n \right] dt \tag{2}$$

on the additional coordinate q^{N+1}. Here c_0 is some arbitrary real constant.

Suppose there to be a differentiable isometric immersion $f : \bar{M} \to M$ so that for each $p \in \bar{M}$ there exists a neighbourhood of $\bar{M}$ whose image is a submanifold of M. The immersion f is used to define what we mean by the $(n-1)$th energy landscape being a nested energy landscape of the nth. At p we have the decomposition

$$T_p M = T_p \bar{M} \oplus (T_p \bar{M})^{\perp} \ , \tag{3}$$

which states that the tangent vector space $T_p M$ can be decomposed into a direct sum of the tangent space $T_p \bar{M}$ and its orthogonal complement $(T_p \bar{M})^{\perp}$. As a consequence, $\bar{M}$ inherits a metric and affine connection $\bar{\nabla}$ from the Eisenhart metric of M. For X and Y vector fields on $\bar{M}$ extended to M it can be shown that $\bar{\nabla}_X Y$ is equal to the component of $\nabla_X Y$ tangential to $\bar{M}$, where ∇ is the affine connection on M [20]. The difference $\nabla - \bar{\nabla}$ uniquely defines the mapping $H : T_p \bar{M} \times T_p \bar{M} \to (T_p \bar{M})^{\perp}$ and the *shape operator* S_Z:

$$\langle S_Z(X), Y \rangle = \langle H(X, Y), Z \rangle \tag{4}$$

along the direction $Z \in (T_p\bar{M})^{\perp}$.

Denote the time of addition of the nth amino acid by t_n, setting $t_n = 0$ and $t_{n+1} = \tau_A$. Let $\gamma_0 \in \bar{M} \subset M$ be the point of the immersed space $\bar{M}$ where the nth amino acid is added to the nascent protein. That particular folding trajectory is then no longer constrained to $\bar{M}$, but continues as a geodesic $\gamma : [0, \tau_A] \to M$ with initial tangent vector $\dot{\gamma}_0 \in (T_{\gamma_0}\bar{M})^{\perp}$ that guarantees γ will leave $\bar{M}$. A fundamental conjecture of the protein folding field is that the energy landscape is shaped such that trajectories originating from different points will converge to a common fold [2–4]. We therefore expect that a similarly constructed trajectory σ, resulting from a delay in addition of the nth amino acid and emanating elsewhere on $\bar{M}$, will converge with γ after a given time interval.

The distance to σ from any point along γ is measured by the Jacobi vector field $J \in T_\gamma M$, which is everywhere orthogonal to the tangent vector field $\dot{\gamma} \in T_\gamma M$ and with suitable initial conditions must satisfy the Jacobi equation

$$\nabla_{\dot{\gamma}} \nabla_{\dot{\gamma}} J = R(\dot{\gamma}, J)\dot{\gamma} \ . \tag{5}$$

Here R is the Riemann curvature tensor on M, whose only non-vanishing components in our chosen coordinate chart are $R_{0i0j} = \partial_i \partial_j V_n$. A small $\|J\|$ implies stability along γ, whereas large $\|J\|$ is indicative of chaotic behaviour [21]; we call a point along γ at which J vanishes and two geodesics converge to a common fold a *focal point* of $\bar{M}$. It is a logical assumption that folding pathways of a good cotranslational folder will converge relatively quickly at each chain length to prevent fluctuations in different τ_A contributing to instability over time. Consequently, the time required for a protein domain of length n to stabilise before addition of the $(n+1)$th amino acid is roughly the interval to the first focal point of $\bar{M}$ along γ.

From Proposition 10.35 in [22] it is possible to derive conditions on $\bar{M}$ and M that guarantee a focal point of $\bar{M}$ over $(0, \tau_A]$. The quadratic form defined by

$$h_{\dot{\gamma}_0}(X) = \langle H(X, X), \dot{\gamma}_0 \rangle \tag{6}$$

for some unit vector $X \in T_{\gamma_0}\bar{M}$ is called the *second fundamental form* of $\bar{M}$ at γ_0 along the direction $\dot{\gamma}_0$ [20]. Provided $h_{\dot{\gamma}_0}(X) \geq 1/\tau_A$ and the *sectional curvatures* of all two-planes containing $\dot{\gamma}$ are positive semidefinite, there is a focal point of $\bar{M}$ on γ before addition of the $(n+1)$th amino acid. This is a powerful result, but the dependence of the conditions

on arbitrary choices of vectors and two-planes makes it difficult to grasp the link with nested energy landscapes. When the dimension N is large it becomes possible to model the sectional curvatures along γ as a stochastic process $\mathcal{K}(t)$. Details on how $\mathcal{K}(t)$ depends on the potential energy V_n can be found in [21, 23], however when the chain length n is small this approximation becomes unsuitable.

We would prefer to derive a more intuitive and exact relationship between V_n, V_{n-1} and the distance to the first focal point of $\bar{M}$. This can be achieved by introducing a new construction on M, but with a cost of ambiguity added to the location of the focal point. It is always possible to pick a hypersurface $P \subset M$ through γ_0 orthogonal to $\dot{\gamma}_0$ so that at γ_0 the shape operator of P agrees with $S_{\dot{\gamma}_0}$. From Warner [24] the first focal point of any such P occurs at least as soon as the first focal point of $\bar{M}$, and we therefore choose P to be the hypersurface whose first focal point occurs furthest along γ. By adapting the proof of Proposition 10.37 in [22] we obtain a set of conditions that must be satisfied if a focal point of $\bar{M}$ is to occur over $(0, \tau_A]$. Provided

$$\frac{1}{N+1}\,\mathrm{trace}(S_{\dot{\gamma}_0}) \geq \frac{1}{\tau_A} \tag{7}$$

and

$$\mathrm{Ric}(\dot{\gamma}, \dot{\gamma}) = \Delta V_n \geq 0 \tag{8}$$

then there can exist a focal point of $\bar{M}$ on γ over the interval $(0, \tau_A]$. However, these conditions do not guarantee existence absolutely. The operator $\mathrm{Ric} : T_pM \times T_pM \to \mathbb{R}$ appearing in 8 is the Ricci tensor on M, whose only non-vanishing components in our chosen coordinate chart are $R_{00} = \Delta V_n$.

The physical interpretations of conditions 7 and 8 follow, and are illustrated graphically in Fig. 1. Expression 8 is a generalisation of the folding funnel hypothesis and states that γ must pass through a subharmonic region or "trough" of the potential V_n if it is to be a stable folding trajectory with which others converge. Trajectories passing over a saddle region of the energy landscape will not converge. Condition 7 is slightly more abstract as it depends on the form of the isometric immersion f and concept of the hypersurface P. Essentially, the expression on the left-hand side of 7 can be thought of as an average of the *negative principle curvatures* of $\bar{M}$ at γ_0, which measures how strongly $(n-1)$th nested energy landscape bends towards $\dot{\gamma}_0$. Intuitively, the larger this quantity is, the smaller the time interval required for two folding trajectories emanating from $\bar{M}$ to converge. The

curvature of $\bar{M}$ required near γ_0 for two geodesics to converge before addition of the $(n+1)$th amino acid is inversely proportionally to the translation rate τ_A. This provides the recursion relation between folding at chain length n and earlier events in the translational pathway that dictate where the point γ_0 appears on $\bar{M}$.

By imposing an even stricter condition on the value of ΔV_n, not necessarily to be satisfied by all energy landscapes, it turns out that we can again guarantee existence of the first focal point over the interval $(0, \tau_A]$. Two points $p, q \in \gamma$ are said to be *conjugate* if a Jacobi field vanishes at both p and q, and so it follows that the first focal point of $\bar{M}$ occurs at least as soon as the first conjugate point along γ (for a more convincing argument see Corollary 2.3 in Warner [24]). From a theorem of Myers [25] we find that if

$$\Delta V_n \geq (N+1)C \ , \tag{9}$$

where $\pi/\sqrt{C} \leq L_\gamma$ (L_γ being the length of γ over $[0, \tau_A]$), then geodesics will converge before addition of the $(n+1)$th amino acid. Condition 9 alone is strong enough to ensure that any two trajectories of suitable length with initial tangent vectors $\dot{\gamma}_0 \in T_{\gamma_0}M$ and $\dot{\sigma}_0 \in T_{\sigma_0}M$ will converge over $(0, \tau_A]$. Satisfying the condition implies the walls of the trough in V_n containing γ are sufficiently steep to draw in all neighbouring trajectories no matter how they originate from the $(n-1)$th nested energy landscape.

In this letter we have provided the geometric foundations of a nested energy landscape theory for cotranslational protein folding. To summarise briefly, the nested energy landscapes for any protein must satisfy certain geometric conditions if the nascent chain is to attain a stable fold on the ribosome. Provided these are met, two folding trajectories leaving the $(n-1)$th landscape at different time points to enter the nth will converge to a common fold before addition of the $(n+1)$th amino acid. The first set of conditions, involving sectional curvatures of the nth landscape and the second fundamental form of the $(n-1)$th, can be re-cast in an averaged, more intuitive version. Firstly, the region of the $(n-1)$th landscape that surrounds the point at which the nth amino acid is added must be sufficiently curved towards the leaving trajectory so as to direct others towards it. The curvature required is inversely proportional to the rate of translation τ_A. A second requirement is that the folding trajectory lies within a trough rather than on a saddle of the nth landscape. In some cases this second condition may be strengthened to ensure convergence of any two trajectories by imposing a lower bound on the steepness of the trough. It is not difficult to

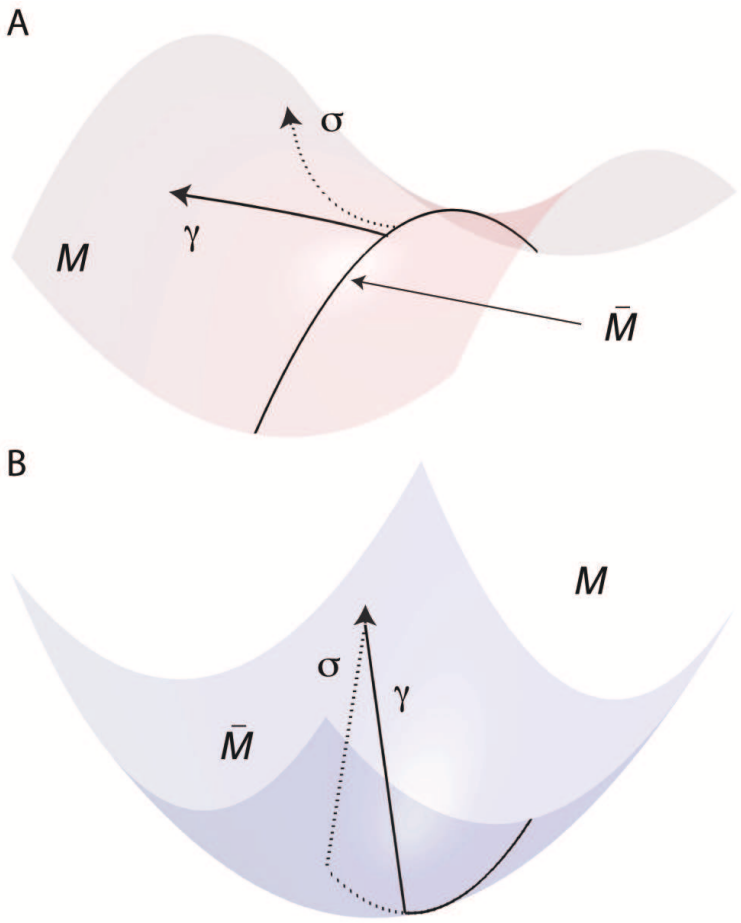

FIG. 1. Simple scheme for an intuitive grasp of conditions 7 and 8. (A) Two trajectories leave the immersed one-dimensional manifold $\bar{M} = \mathbb{R}$ at different time points to enter the surface M. Since M is hyperbolic, the two trajectories will not converge. (B) In this case $\bar{M}$ is a two-dimensional surface immersed in $M = \mathbb{R}^3$. Trajectories leave $\bar{M}$ at different time points, but curvatures of the immersed surface and ambient manifold are sufficient to allow convergence of γ and σ over the interval $[0, \tau_A]$.

envisage the obvious extension of these results to more general scenarios involving growth of atomic clusters or elongation of generic polymers.

This work was made possible by the award of an MRC studentship to the author.

---

* davidt@mrc-lmb.cam.ac.uk

[1] T.M. Schmeing and V. Ramakrishnan, Nature **461**, 1234 (2009).

[2] P.E. Leopold, M. Montal, and J.N. Onuchic, Proc. Nat. Acad. Sci. USA **89**, 8721 (1992).

[3] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes, Proteins **21**, 167 (1995).

[4] J.N. Onuchic, P.G. Wolynes, Z. Luthey-Schulten, and N.D. Socci, Proc. Nat. Acad. Sci. USA **92**, 3626 (1995).

[5] A.N. Fedorov and T.O. Baldwin, J. Biol. Chem. **272**, 32715 (1997).

[6] L.D. Cabrita, C.M. Dobson, and J. Christodoulou, Curr. Opin. Struct. Biol. **20**, 33 (2010).

[7] K.G. Ugrinov and P.L. Clark, Biophys. J. **98**, 1312 (2010).

[8] W.J. Netzer and F.U. Hartl, Nature **388**, 343 (1997).

[9] A.V. Nicola, W. Chen, and A. Helenius, Nat. Cell Biol. **1**, 341 (1999).

[10] P.L. Clark and J. King, J. Biol. Chem. **276**, 25411 (2001).

[11] A.H. Elcock, PLoS Comput. Biol. **2**, e98 (2006).

[12] A.A. Komar, T. Lesnik, and C. Reiss, FEBS Lett. **462**, 387 (1999).

[13] C.J. Tsai *et al.*, J. Mol. Biol. **383**, 281 (2008).

[14] G. Zhang, M. Hubalewska, and Z. Ignatova, Nat. Struct. Mol. Biol. **16**, 274 (2009).

[15] E.P. O'Brien, M. Vendruscolo, and C.M. Dobson, Nat. Commun. **3**, 868 (2012).

[16] P. Ciryam, R.I. Morimoto, M. Vendruscolo, C.M. Dobson, and E.P. O'Brien, Proc. Nat. Acad. Sci. USA **110**, 396 (2013).

[17] LN. Mazzoni and L. Casetti, Phys. Rev. Lett. **97**, 218104 (2006).

[18] L.N. Mazzoni and L. Casetti, Phys. Rev. E **77**, 051917 (2008).

[19] L.P. Eisenhart, Ann. Math. **30**, 591 (1928).

[20] M. P. do Carmo, in *Riemannian Geometry* (Birkhäuser, 1992), Vol. 1, p. 124.

[21] L. Casetti, M. Pettini, and E.G.D. Cohen, Phys. Rep. **337**, 237 (2000).

[22] B. O'Neill, in *Semi-Riemannian Geometry With Applications to Relativity*, Pure and Applied Mathematics Vol. 103 (Academic Press, 1983), p. 263.

[23] L. Casetti, C. Clementi, and M. Pettini, Phys. Rev. E **54**, 5969 (1996).

[24] F.W. Warner, Trans. Amer. Math. Soc. **122**, 341 (1966).

[25] S. B. Myers, Duke Math. J. **8**, 401 (1941).